DOCUMENT RESUME

ED 298 174 TM 012 323

AUTHOR Marso, Ronald N.; Pigge, Fred L.
TITLE An Analysis of Teacher-Made Tests: Testing Practices, Cognitive Demands, and Item Construction Errors.
PUB DATE Apr 88
NOTE 50p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classroom Techniques; Cognitive Measurement; Elementary Secondary Education; *Public School Teachers; *Teacher Made Tests; *Test Construction; Test Format; Testing; Test Items

ABSTRACT

This study encompassed the collection of teacher reported (N=326) testing practices and the direct assessment of teacher-made tests (N=175) for item cognitive functioning levels and construction errors. Focus was on assessing the nature and quality of teacher-made tests used in public school classrooms and describing the classroom teachers' testing preferences. It was found that the classroom teachers prepared and administered many formal teacher-made tests during the school year (X=54.1); they wrote most of their own test items; they most frequently used multiple-choice, matching, and short response items but infrequently used essay items; they infrequently completed post-hoc statistical analyses of their tests; most teachers' tests and test items functioned primarily at the knowledge cognitive level with the exception of the math tests; and matching exercises followed by completion and essay items contained the most construction errors per exercise. Item cognitive functioning levels, testing practices, and item construction error frequencies significantly differed when the tests and teacher survey responses were classified by grade level and by subject area, but only cognitive functioning levels differed by the amount of teaching experience or school setting (urban, rural, and suburban). Seven tables are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298174

An Analysis of Teacher-made Tests: Testing Practices, Cognitive Demands, and Item Construction Errors

Ronald N. Marso and Fred L. Pigge

College of Education and Allied Professions

Bowling Green State University

Bowling Green, Ohio 43403

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RONALD N. MARSO & FRED L. PIGGE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the annual meeting of the

National Council on Measurement in Education

New Orleans, Louisiana

April 6-8, 1988

Running Head: ANALYSIS TEACHER-MADE TESTS

TM012 323

BEST COPY AVAILABLE

Abstract

This study encompassed the collection of teacher reported (N = 326) testing practices and the direct assessment of teacher-made tests (N = 175) for item cognitive functioning levels and construction errors. It was found that the classroom teachers prepared and administered many formal teacher-made tests during the school year ($\bar{X}$ = 54.1); wrote most of their own test items; most frequently used multiple-choice, matching, and short response items but infrequently used essay items; infrequently completed post-hoc statistical analyses of their tests; most teachers' tests and test items functioned primarily at the knowledge cognitive level with the exception of the math tests; matching exercises followed by completion and essay items contained the most construction errors per exercise; and item cognitive functioning levels, testing practices, and item construction error frequencies significantly differed when the tests and teacher survey responses were classified by grade level and by subject area but only cognitive functioning levels differed by amount of teaching experience or school setting (urban, rural, and suburban).

An Analysis of Teacher-made Tests: Testing Practices, Cognitive Demands, and Item Construction Errors

It is rather widely accepted that the day-to-day impact of teacher-made testing has more influence upon what happens in the typical classroom than do standardized tests; yet, far less is known about the nature of teacher-made tests and related testing practices (Fleming & Chambers, 1983; Stiggins, 1985). For example, Gullickson (1984) described existing research on teacher-made testing practices as limited and idiosyncratic, and Dwyer (1982) stated that the advice given to preservice and inservice teachers regarding the use of teacher-made tests reflects a consensus of professional judgement rather than a foundation of empirical research. The spareness of this research is further limited by the narrow scope of the research methods used. Most existing research on teacher-made tests and testing practices has been based on self-report procedures. And perhaps even further impeding the effective classroom use of teacher-made tests, some research suggests that university tests and measurement courses may not be meeting the needs of the classroom as identified by inservice teachers (Gullickson, 1986; Gullickson & Ellwein, 1985; Stiggins & Bridgeford, 1985) and that many teacher-training institutions do not require their

preservice teachers to take a course in tests and measurements (Lambert, 1980-81).

Relative to teacher-made testing practices, Gullickson (1984) reported that most teachers feel that frequent testing is desirable, that teachers perceive students as desiring frequent tests, and that most teachers test at least once every two weeks in most subject areas. From a survey of 228 teachers, Stiggins and Bridgeford (1985) reported that teachers in the upper grades rely more heavily on teacher-made tests rather than on standardized or publisher-made tests, that types of assessments used varied by subject area, and that about three-fourths of the teachers in their sample expressed a desire to improve their teacher-made tests. And Rogers (1985), using interview procedures, reported that some teachers rely solely on publisher-made tests, that some teachers rely solely on teacher-made tests, but that most teachers used both types of tests. He also reported that most teachers relied more on paper-pencil tests than upon less formal assessment procedures, but observations and ratings of products or behaviors were considered by teachers to be desirable supplements to the paper and pencil tests.

Relatively few reports of direct assessments of samples of teacher-made tests have appeared in the research literature; consequently, little is known about the essential characteristics

of teacher-made tests such as typical item cognitive demand levels or the types of item construction errors most commonly found on these tests. Gullickson (1984), in his previously reported survey study, reported that his sample of teachers felt that their teacher-made tests did not effectively assess student ability to apply what they had learned, but the teachers' tests were not directly examined to support or refute this contention. Fleming and Chambers (1983) did report the results of an extensive assessment of 342 tests developed by teachers in grades one through 12. These researchers used Bloom's (1956) taxonomy of cognitive levels to classify the 8,800 test items on these tests, and they also examined each test for the presence of various item or test format construction errors. They found that short response items, including fill-in-the-blank items, were most frequently used followed in popularity by matching exercises. True-false and essay type items were least frequently used in this sample of tests. The junior high level teachers asked the most knowledge level questions with 94% of their items judged to be functioning at this level; whereas the elementary and senior high teachers' tests were determined to have 69% of their items functioning at the knowledge level. In analyzing the tests by subject area, however, these researchers found that only in the math and science tests were the items judged to be functioning at the upper range of cognitive levels with

predominantly knowledge level items found in all other subject areas. Their assessment of the tests for construction errors revealed that: a) directions were absent in approximately one-third of the tests, b) test items were not numbered in approximately one-half of the elementary grade level tests, c) many of the tests at the junior and senior high school levels did not have items numbered consecutively throughout the test, d) approximately 15 to 20 percent of the tests exhibited grammatical, spelling, or punctuation errors, e) many one or two word stem multiple-choice items were found, f) a large proportion of the tests were found to be illegible and many were handwritten, and g) the short response type items tended to be ambiguous and allowed more than one correct answer.

Two additional but less extensive analyses of actual samples of teacher-made tests were also located in the professional literature. Bloom's taxonomy of cognitive levels was used to classify the teacher constructed test items in both of these investigations. Billeh (1974) reported an analysis of 33 science tests constructed by seventh through tenth grade teachers in the Beirut, Lebanon schools. He found that 72% of the test items measured at the knowledge level, 21% at the comprehension level, 7% at the application level, and no items were found to measure at the analysis, synthesis, or evaluation levels. Additionally, he found that the cognitive levels of the test items did not vary

by the grade level of the teacher-made tests or by the extent of the training of the teachers constructing the tests; however he did find that the cognitive levels of the test items differed by science subject and by the amount of teaching experience of the teachers. The more experienced teachers used more knowledge level items, and the physics teachers used fewer knowledge level items than either the biology or the chemistry teachers. Black (1980) also reported an analysis of teacher-made science tests. These science tests were constructed by 48 secondary teachers in Nigeria. He also found that the cognitive demand levels of the test items varied between the science subjects, that none of these items measured beyond the application level, and that the extent of teacher training did not appear to influence the cognitive functioning levels of their tests. The proportions of items found to be functioning at the various cognitive levels for the various subject area tests were: biology 94% knowledge and 6% comprehension; chemistry 66% knowledge, 26% comprehension, and 8% application; and for physics 56% knowledge, 18% comprehension, and 26% application.

Purpose

The basic purpose of this study was to assess the nature and the quality of teacher-made tests being used in public school classrooms through an analysis of a sample of these teacher-made

tests. A secondary purpose of the study was to describe the classroom teachers' testing preferences and practices such as use of post-hoc test statistical procedures, frequency of tests scheduled, the proportion of the items on their teacher-made tests that they write themselves, and the extent of their use of various item types. More specifically, the following six hypotheses were stated to guide this study:

1) The types of items most frequently found on the teacher-made tests will not differ significantly by the teachers': a) grade level, b) subject area, c) school setting, or d) years of teaching experience.
2) The cognitive levels of the test items found on the teacher-made tests will not differ significantly by the teachers': a) grade level, b) subject area, c) school setting, or d) years of teaching experience.
3) The reported number of formal teacher-made tests given in a typical school year will not differ significantly by the teachers': a) grade level, b) subject area, c) school setting, or d) years of teaching experience.
4) The reported use of post hoc test statistical analysis will not differ significantly by the teachers': a) grade level, b) subject area, c) school setting, or d) years of teaching experience.

5) The reported proportion of test items used on teacher-made tests which were constructed by the teachers themselves will not differ signigicantly by the teachers': a) grade level, b) subject area, c) school setting, or d) years of teaching experience.

6) The frequencies of test item and test format errors found on the teacher-made tests will not differ significantly by: a) grade level, b) subject area, c) school setting, or d) years of teaching experience.

Method

From a Spring 1986 state-wide assessment of teacher testing and evaluation competencies completed by a stratified random sample of 580 supervisors and principals and by 326 former Bowling Green State University students who had graduated during the 1975-1986 period and who were teaching full-time in Ohio during the 1985-86 school year, a sample of 175 teacher-made tests were collected to allow a direct analysis of teacher testing practices and proficiencies. The selected classroom teachers were asked to provide a copy of their most recently administered teacher-made test for a subject other than spelling or math (unless they were teaching secondary mathematics). These teachers were also asked to answer a set of questions regarding their testing preferences and practices and a set of demographic-type questions regarding themselves and their

employing school. Only teachers assigned to regular elementary or secondary level classrooms were asked to participate in this component of the study.

Survey instrument

In the demographic section of the survey form each teacher was asked to report his/her teaching grade-level assignment, subject area of specialization if a secondary teacher, type of employing school (rural, urban, or suburban), and the number of years she/he had taught. In the testing practices and preferences section of the survey instrument the teachers were asked to report how frequently they calculated means and standard deviations, estimated the reliability, and completed item analysis after having administered their teacher-made tests. The response continuum for these survey items ranged from never (1) to always (5). Additionally the teachers were asked to indicate: the proportion of the questions on their formal teacher-made tests that had been written by themselves on a scale from very few (1) to almost all (5), the number of formal tests excluding spelling or other quizzes they gave in a typical subject or class on a scale from one or more each week (1) to two or fewer per semester (5), the number of formal tests they gave during a typical school year including all classes, and of all test questions they used in a typical school year the approximate percentage of this total that were of the completion, matching,

true-false, multiple-choice, essay, problems, or other item types.

The teachers responding to the survey instrument consisted of 122 elementary, 191 secondary, and 13 specialized area (certified K-12) teachers. One hundred and thirty-four (134) described their employing schools to be rural, 150 suburban, and 42 urban. When asked to report the number of years of full-time teaching experience, 68 of the teachers reported one to three years, 85 reported four to six years and 173 of the teachers reported having seven or more years of teaching experience. The subject area classifications of the secondary teachers were: 45 business education, 33 science, 41 mathematics, 30 English, 32 social studies, and 10 other areas of specialization. Each of these teachers had completed an undergraduate tests and measurements course taught by one of approximately ten professors providing instruction for the course during the 1975-1985 time period.

Of the 326 teachers returning usable survey forms, 175 (54%) also enclosed a copy of their most recent formal teacher-made test. These tests, regardless of grade level, when classified by subject area consisted of 30 history/social studies, 36 science, 29 business education, 32 mathematics, 28 English, and 20 tests within nine other specializations with insufficient numbers to be included in distinct subject area categories.

Sample of teacher-made tests

The sample of 175 teacher-made tests included a total of 6504 test items and 455 item exercises. The test items within the sample of tests were classified independently by two judges using Bloom's taxonomy of six cognitive demand levels (knowledge, comprehension, application, analysis, synthesis, and evaluation). If the judges differed in their classification of an item or exercise, the item or exercise was reexamined until a consensus was reached. Each test and each test exercise was also examined for format and item construction errors. A test exercise was defined for this study as a group of items of a similar item type, and item construction error criteria were selected from a review of several test construction texts designed for preservice education courses. A total of eight item type classifications (completion, essay, multiple-choice, etc.), 10 item format construction error criteria (does the test have complete directions? are item types grouped together? are the items numbered consecutively? etc.), and 66 item construction error criteria (incomplete stems, implausible alternates, specific determiners, etc.) were identified from these procedures and used in the assessment of the sample of teacher-made tests. An item construction error, if present, was recorded once per item exercise rather than for each time that particular error type may have occurred within the item exercise. In other words whether

or not a construction error appeared only on one item or on several items within the same item exercise a tally of '1' was recorded for that particular error in order to provide a stable base of comparison across tests which varied in their number of test items.

Data collection and analysis

The percentage of teachers responding to each of the testing preferences and practices survey items were calculated. Each of the individual test items was classified according to the six cognitive demand levels described by Bloom, the number of test item types per test and the total number of exercises in the total sample of tests were tallied, and each item exercise and each test was examined for test construction errors with the frequency of each tallied.

The teacher responses to the seven testing practices and preferences items were analyzed using one-way ANOVA procedures on the "scores" produced by each of the seven items. More specifically the dependent variables for these analyses were the teacher responses on each of the five-point response scales (first five items), the reported number of "major" tests given in a typical school year (the sixth item), and the relative percentage of each identified test item type making up the teachers' cumulative yearly efforts at constructing tests (the last item in this survey section). The classification

(independent) variables in these ANOVA analyses were: a) school setting (rural, suburban, or urban), b) teaching grade level assignment (elementary or senior high school), c) if high school teachers, their subject area specialization, and d) years of teaching experience (1-3, 4-6, and 7 or more years). The seven items as they appeared on the survey form with a summary of teacher responses are reported on Table 1.

The frequency data obtained from the direct assessment of the tests and the items or item exercises within these tests were analyzed by chi-square procedures. The four classification (independent) variables used in the analysis of the teacher responses to the survey items were also used in these analyses while the assessment frequency scores were used as the dependent variable. For example these latter "scores" were made up of the frequency of a construction error type, the number of completion items used, or the number of knowledge level items used.

Results

Teachers' testing preferences and practices

Most of the teachers reported infrequent use of statistical procedures following the administration of their teacher-made tests: 80% of the responding teachers indicated that they never or rarely calculated test means and standard deviations (5% responded always or nearly always), 60% indicated that they never or rarely estimated the reliability of their tests (15% responded

always or nearly always), and 54% of the responding teachers indicated that they never or rarely completed item analyses of their tests (16% nearly always or always).

The teachers did report frequent scheduling of formal tests (excluding quizzes and spelling tests) in a typical class and in a typical school year. The mean number of teacher-made tests administered during a typical school year was 54.1 with 31% of the teachers reporting the administration of 60 or more formal tests and 15% of the teachers reporting the administration of 100 or more formal tests in a typical school year. When asked how frequently they scheduled formal tests in a typical class, 20% reported scheduling one or more formal tests each week, 49% reported one every two weeks, 15% one per month, 7% three or four per semester, and only 6% reported scheduling two or fewer formal tests in a typical class during a school semester.

Over one-half of the teachers reported writing three-fourths or nearly all of the items used on their teacher-made tests. More specifically, approximately 37% of the teachers reported writing almost all of their test questions, 20% about three-fourths of their items, 19% about one-half, 8% about one-fourth, and 14% reported writing very few of the test items used in assessing the progress of their students. For all the test items used during an entire school year, the teachers were asked to estimate the proportion of each item type used; the average of

their percentage responses for each item type were: 23% problems, 19% multiple-choice, 16% completion, 16% essay, 14% matching, and 12% true-false. These survey items and teacher responses to them are presented on Table 1.

- - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - -

When the teacher responses to the seven testing practices and preferences items were classified by teacher and school characteristics, it was found that neither the school setting (rural, urban, and suburban) nor the years of teaching experience (1-3, 4-6, and 7 or more years) classifications revealed mean differences; whereas, the grade level and subject area classifications of the teacher responses each revealed mean differences on five of the survey items. As shown on Table 2, the secondary teachers as compared to the elementary teachers indicated that they: more frequently calculated means and standard deviations for their tests (item 1.a, elem. $\bar{X}$ = 1.58, 2cdary $\bar{X}$ = 1.89, F = 8.67, p = .01), more frequently completed item analysis procedures (item 1.c, elem. $\bar{X}$ = 2.20, 2cdary $\bar{X}$ = 2.46, F = 3.84, p = .05), wrote proportionately more of their own test items (item 2, elem. $\bar{X}$ = 2.66, 2cdary $\bar{X}$ = 4.12, F = 96.87, p = .001), and gave more frequent tests during a typical course (item 1.d., elem. $\bar{X}$ = 2.45, 2cdary $\bar{X}$ = 2.14, F = 6.51, p = .01).

Mean differences between the elementary and secondary teachers for these two testing practices were not statistically significant: frequency of calculating reliability after administering teacher-made tests and the number of formal tests given in a typical school year.

- - - - - - - - - - - - -

Insert Table 2 about here

- - - - - - - - - - - - -

Additionally and as presented in Table 3, the secondary teachers as compared to the elementary teachers reported using proportionately more essay items (elem. $\bar{X}$ = 7.33, 2cdary $\bar{X}$ = 13.31, F = 10.06, p = .002) and more problem type items (elem. $\bar{X}$ = 13.98, 2cdary $\bar{X}$ = 26.33, F = 12.59, p = .001), but somewhat fewer completion (elem. $\bar{X}$ = 18.97, 2cdary $\bar{X}$ = 15.33, F = 3.00, p = .08), and fewer multiple-choice item types (elem. $\bar{X}$ = 24.48, 2cdary $\bar{X}$ = 16.72, F = 11.21, p = .001) during a typical academic year. The elementary and the secondary teachers did not differ significantly in their reported use of matching, true/false, and "other" item types.

The subject area classification of teacher responses to the testing practice or preference items revealed (see bottom section of Table 2) that social studies teachers reported less frequent calculation of test means or standard deviations than did the science teachers, but neither of the means of these two groups of

teachers differed from the means of the other three teacher specialization groups (science $\bar{X}$ = 2.39, social studies $\bar{X}$ = 1.46, English $\bar{X}$ = 1.73, math $\bar{X}$ = 1.95, business $\bar{X}$ = 1.86, F = 4.18, p = .01) with the post-hoc mean pair comparisons set at the .10 level of significance. Similarly, the social studies teachers as compared to the business teachers reported less frequent use of item analysis techniques (business $\bar{X}$ = 2.84, math $\bar{X}$ = 2.58, science $\bar{X}$ = 2.39, English $\bar{X}$ = 2.17, social studies $\bar{X}$ = 2.09, F = 2.99, p = .02) but wrote more of their own test items (social studies $\bar{X}$ = 4.50, science $\bar{X}$ = 4.33, English $\bar{X}$ = 4.13, math $\bar{X}$ = 4.05, business $\bar{X}$ = 3.67, F = 3.18, p = .02) than did the business teachers; whereas neither the means of the social studies nor the means of the business groups differed significantly from the means of the three other groups of teachers on these two items. Additionally, the English teachers reported using fewer formal tests during a typical course than did any of the other groups of teachers (English $\bar{X}$ = 2.77, math $\bar{X}$ = 2.10, business $\bar{X}$ = 2.09, science $\bar{X}$ = 2.00, social studies $\bar{X}$ = 1.71, F = 6.58, p = .001); no other pair-wise mean differences were significant ("scores" for this survey item were: 1 = one or more each week through 5 = two or fewer per semester, thus lower means indicate more frequent administration of teacher-made tests).

When the proportionate use of each item type relative to total number of test items used in preparing tests over a school

year was examined relative to the teachers' subject area classification, significant differences among the five specializations were noted on each of the test item types as reported on Table 3. Social studies teachers reported using more completion type items than math teachers with neither of these means being significantly different from the means of the other three groups (social studies $\bar{X}$ = 21.97, science $\bar{X}$ = 16.58, business $\bar{X}$ = 15.44, English $\bar{X}$ = 13.90, math $\bar{X}$ = 7.66, F = 3.54, p = .008). For matching exercises the math teachers reported less use of this item type than did each of the other four groups of teachers (science $\bar{X}$ = 20.33, social studies $\bar{X}$ = 19.56, English $\bar{X}$ = 15.57, business $\bar{X}$ = 14.38, math $\bar{X}$ = 3.41, F = 11.29, p = .001). The math teachers also reported less use of the true-false item type than did the social studies and business education teachers, the English teachers reported less use of true-false items than did the business education and social studies teachers, and the science teachers reported less use of the true-false items than did the business education teachers (business $\bar{X}$ = 14.69, social studies $\bar{X}$ = 8.52, English $\bar{X}$ = 7.20, math $\bar{X}$ = 3.44, F = 12.10, p = .001). The math teachers also reported less use of the multiple-choice type items than did each of the other four groups with none of the other four group means differing significantly one from another (English $\bar{X}$ = 23.63, science $\bar{X}$ = 26.36, social studies $\bar{X}$ = 18.75,

business $\bar{X}$ = 17.00, math $\bar{X}$ = 3.17, F = 13.10, p = .001). The social studies and English teachers reported greater use of the essay item type than did the other three groups of teachers (English $\bar{X}$ = 29.87, social studies $\bar{X}$ = 21.06, business $\bar{X}$ = 7.44, science $\bar{X}$ = 9.85, math $\bar{X}$ = .32, F = 21.93, p = .001). Last and as expected, the math teachers reported more use of problem type items than did each of the other four groups of teachers; whereas the business education and the science teachers did not differ in their reported use of problem type items they reported less use than did the math teachers but indicated a greater use of these items than did the English and social studies teachers (math $\bar{X}$ = 78.76, business $\bar{X}$ = 26.47, science $\bar{X}$ = 15.48, social studies $\bar{X}$ = 1.25, English $\bar{X}$ = 1.17, F = 106.55, p = .001).

- - - - - - - - - - - - - -

Insert Table 3 about here

- - - - - - - - - - - - - -

Assessment of the teacher-made tests

Item types used. When the tests were assessed by item type, a total of 455 test exercises were identified among the 6529 items contained on the 175 teacher-made tests. The number of items by type from highest to lowest found on this sample of teacher-made tests with the percentage of this number to the total 6529 items were: multiple-choice 1317 (20%), matching 1261 (19%), short response 1093 (17%), true-false 935 (14%), problems

896 (14%), completion 549 (8%), interpretive exercise 362 (6%), essay 64 (1%), and unclassified items not fitting any of the preceding categories 52 (1%). Examining item use by exercise (a group of items of the same type) rather than by total frequency of the item leads to a somewhat different view of the characteristics of teacher-made tests. The most frequently appearing item exercises in terms of the number of teacher-made tests that they were found on and the percentage of the total number of the exercises of this type (the number of exercises divided by 175 representing the total number of tests) were: short response 89 (51%), matching 78 (45%), true-false 69 (39%), multiple-choice 65 (37%), problems 54 (31%), completion 48 (27%), interpretive exercises 30 (17%), essay 22 (13%), and unclassified exercises 6 (3%). In other words, considering the total number of individual items, more multiple-choice items (1317) were found on the teacher-made tests than any other item type; however, as groups of items (item exercises) the short response (89), the matching (78), and the true-false (69) exercises were more frequently used than the multiple-choice exercise (65).

Relative to the number of questions found on each test, the average length of the tests was found to be 37.9 items with a range from three to 125 items and a standard deviation of 23.6. Only 26% of the tests contained fewer than 20 items, 56% contained 35 or fewer items, and 78% contained 50 or fewer items.

The comparisons of the frequencies of the use of various item type exercises by teachers' school setting, grade level, subject field, and years of teaching experience revealed that teachers when classified other than by subject fields used similar item type exercises on their teacher-made tests. Only three significant differences were found in use of item types when the teacher-made tests were analyzed by school setting, grade level, and years of teaching experience. The school setting and the teaching experience classifications revealed a significant difference in the number of exercises used for only one item type. Teachers with 1-3 years of experience used more interpretive items (found on 30% of their tests) as compared to the more experienced teachers (found on 15% of the teacher tests with 4-6 years and on 12% of those with 7-10 years), $x^2 = 6.68$, $p = .04$ (goodness of fit chi-square using frequencies reported); and suburban teachers more frequently used problem exercises (found on 71% of their tests as compared to urban 36% or rural teachers 30%), $x^2 = 27.93$, $p = .001$. The elementary and secondary teachers differed in the frequency of their use of matching and problem exercises. Elementary teachers used more matching exercises (59%) but fewer problem exercises (5%) as compared to the secondary teachers with matching exercises found on 40% of their tests ($x^2 = 4.64$, $p = .03$) and with problem exercises found on 37% of their tests ($x^2 = 13.74$, $p = .001$).

Conversely and as indicated on the top part of Table 4, the subject field classification of the use of the various test exercises revealed significant differences in the use of all item type exercises with the single exception of the short response item exercise. Less frequent use of all item types except the problem type were noted on the math tests as compared to the other subject fields; 97% of the math tests had problem exercises. English teachers used more matching (75%) and essay (32%) exercises than any other fields; whereas the science (6%), business (10%), and math (0%) fields all made very infrequent use of the essay items. In addition business teachers made relatively less frequent use of multiple-choice (28%), and social studies teachers used relatively fewer interpretive exercises (10%) compared to the teachers in the other subject fields.

- - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - -

Item cognitive levels. The two judges reached consensus on the classification of 6504 (of the total 6529 items) items with 72% of these being classified as functioning at the knowledge level, 11% at the comprehension level, 15% at the application level, 1% at the analysis level, and fewer than 1% of the items classified as functioning at the synthesis and evaluation levels. The above percentages of the items functioning at the various

cognitive levels relative to the total 6504 items appear to be relatively acceptable until analyzed test by test. When reexamining the items by individual test, it was found that most tests consisted of items functioning exclusively or predominately at the knowledge level. Nearly all of the higher cognitive level items were located on the mathematics and science tests. Only the teacher-made problem type items were found to be consistently functioning beyond the knowledge level; consequently the math and science tests accounted for 857 (47%) of the total 1834 items classified as functioning beyond the knowledge level as shown in section A of Table 5. The percentage of items measuring beyond the knowledge level by item type was found to be: problem 96%, essay 53%, unclassified 46%, interpretive exercises 35%, short response 24%, true-false 20%, multiple-choice 15%, matching 8%, and completion 2%.

- - - - - - - - - - - - - -

Insert Table 5 about here

- - - - - - - - - - - - - -

When the cognitive functioning levels of the items from the teacher-made tests were examined within the teacher classifications of school setting, grade level, years of teaching experience, and subject field, it was found that the less experienced teachers constructed somewhat more comprehension level test items as compared to all test items constructed than

did the more experienced teachers (1-3 years [17%], 4-6 years [12%], and 7-10 years [10%]), $x^2 = 57.16$, $p = .001$ (chi-square based on frequencies from a 3x3 contingency table); that the teachers employed by rural schools constructed somewhat fewer knowledge level items (and concomitantly somewhat more comprehension items) as compared to all items constructed than did the teachers employed by urban and suburban schools (rural [69%], urban [74%], and suburban [74%], $x^2 = 31.08$, $p = .001$; that elementary grade teachers constructed somewhat more knowledge level items (76%) and comprehension level items (18%) but fewer items functioning at the higher levels (6%) as compared to the secondary teachers (knowledge items [71%], comprehension [12%], and higher levels [17%]), $x^2 = 111.05$, $p = .001$; and that the cognitive functioning levels of items constructed by teachers in the five subject field classifications differed among all possible pairings. The social studies tests were found to have the highest percentage of knowledge level items to total items constructed at 98% (science [80%], English [77%], business [79%], and math [7%]), with English having proportionately the most comprehension level items at 21% (math [14%], business [14%], science [11%], and social studies [29%]), and with math having proportionately the most higher cognitive level items to all items constructed at 79% (science [9%], business [7%], English [2%], and social studies [0%]. The results of these 10 2x3

chi-square tests of independence are presented in section B of Table 5.

Test format and item construction errors. The analysis of the 455 item exercises for item construction errors and the analysis of the 175 tests for item format errors resulted in the identification of 853 item construction errors and 281 test format errors as summarized on Table 6. Construction errors were most frequently found in matching exercises with an average of 6.4 different types of errors in each exercise followed by completion exercises with an average of 2.2, and essay exercises with an average of 1.5 different types of errors per exercise. Construction errors were least frequently found in the interpretive exercises with an average of 0.2 different types of errors in each exercise followed by problem exercises with an average of 0.5, short response exercises with an average of 0.7, multiple-choice exercises with an average of 0.8, and true-false exercises with an average of 1.0.

As the data on Table 6 indicate, a total of 281 test format errors were identified on the 175 tests. Most frequent format errors were absence of directions (29% of all errors), answering procedures not clear (22% of all errors), and items not consecutively numbered throughout the test (17% of all errors). This data for all 10 test format criterion are presented in Table 6 section B.

When the test item and test format construction errors were examined within the four subject classifications of school setting, grade level, subject field, and years of teaching experience, it was found that the average number of construct type errors per test exercise and the average number of test format type errors per test did not differ significantly when the tests were classified by the teachers' school setting or by the years of teachers' teaching experience. Further, the grade level classification revealed significant differences only for multiple-choice exercises with fewer average item construction type errors found on the multiple-choice exercises for the elementary teachers as compared to the secondary teachers, $x^2 = 5.33$, $p = .02$. Conversely, the subject field classification of the tests revealed differences for the short response and true-false exercises and for test format construction errors. Fewer test format construction type errors per test were found on the math tests as compared to the other tests (math 8, business 12, science 16, English 18, and social studies 25), $x^2 = 10.43$, $p = .03$; fewer construction errors were found per exercise on the short response exercise in the math tests (math 0, English 5, business 6, science 7, and social studies 13), $x^2 = 14.00$, $p = .01$; and fewer construction errors per true-false exercise were found on the English (4), and science (5), tests as compared to business (17) and social

studies (12), $x^2 = 11.89$, $p = .01$ (math was excluded from this comparison as fewer than five short response exercises were available for analysis).

- - - - - - - - - - - - -

Insert Table 6 about here

- - - - - - - - - - - - -

A specific listing of the item construction errors by item type is presented on Table 7. The most common types of construction error identified for the completion items were as follows with the percentage of total errors for this type of exercise noted: questions not complete interrogative sentence (30%), blanks placed in the middle of the statement rather than to the left or to the right (29%), the questions appeared to be statements taken from a textbook rather than stated questions (17%), and the questions were constructed with more than a single answer called for (e.g. more than a single blank per question). Similarly the most common types of construction error found on the true-false or alternate response exercises were: student required to write out answers rather than to circle T or F or simply place T or F in answer space which does not make efficient use of testing or scoring time (28%), statements contain more than a single idea resulting in a true or false for different ideas in a single question (23%), questions stated in a negative form rather than restated in a positive form with the key changed

(21%), and the presence of a specific determiner (always, never, etc.) acting as a clue (11%).

- - - - - - - - - - - - - -

Insert Table 7 about here

- - - - - - - - - - - - - -

The most common type of construction errors identified on the matching exercises were as follows with the percentage of total errors for this type of exercise noted: the premise and response columns were not titled (14%), directions allowed the elimination of responses (14%), the response column was not ordered alphabetically or chronologically when they should have been (12%), the directions for the exercise did not exist or did not spell out the basis for the match (11%), and the answering procedures were not specified (e.g., draw lines between, place the letter before the correct response, write out correct response to left of premise, etc.). Similarly, the most common errors found on the multiple choice exercises were: the alternates were not placed in column (either one or two) or row but placed in narrative-paragraph form (40%), incomplete stems (23%), negative words were not underlined or capitalized (17%), and the all above or none of above alternate was not appropriately used or was used as a "filler" alternate throughout an exercise (9%).

The most common construction errors found on the essay exercises with the percentage of the numbers of this type of error relative to all errors found on this item type for the essay exercises were: the response or answer expectations were not clear (e.g., "list the reasons for the Civil War," not clear as to how many reasons were to be listed and as to if these reasons were also to be explained) which accounted for 41% of the errors found, scoring points assigned unrealistically high (e.g., 15 to 20 points for a relatively simple, single paragraph question being weighted more heavily than complete exercises of matching, true-false, etc. on the same test) which accounted for 21% of the errors found, optional questions provided (15%), and restricted questions were not provided for limited time and space (e.g., "explain the causes of the Civil War" for which the topic is so broad that books are written rather than "explain how the following three factors led to the Civil War...") which accounted for 9% of the total number of errors found on these exercises. Similarly, the most frequent construction errors found on the problem exercises were: the test included only calculations with no other item types to sample student understanding of concepts (77%), the test did not provide a range of easy to difficult problems to assess "process" as well as "answer accuracy" (12%), and degree of accuracy was not denoted where it appeared to be necessary (e.g., the problem presented units in feet and inches

but it was not clear whether the answer should be in both or one or the other) accounting for 8% of the total number of errors found on this type of exercises.

Only one type of construction error, lack of an objective response format, was noted on the interpretive exercises. This error was found on six of the 30 interpretive exercises. Three assessment criterion were used for the short response exercises resulting in the following percentage to total exercises errors: item requests just a simple "listing" knowledge level response (84%), question is ambiguous or response expectations unspecified (e.g., "Who was George Washington?" Would any of the following be accepted? Our first president, he could not lie about chopping down a cherry tree, etc.) which accounted for 11% of the errors identified for this type of item, and unrealistically high score values assigned (e.g., five points for the recall of a one-phrase response) which accounted for the remaining 5% of the errors for this type of exercise.

Summary, Discussion, and Implications

The data collected from the assessment of the sample of teacher-made tests and from the survey of teacher testing practices led to the rejection of the six stated null hypotheses. It was found that the reported use of various test item types, the reported frequency of tests scheduled during a typical class (but not the total number of tests given by a teacher during an

academic school year), the reported proportion of test items used in testing written by the teachers themselves, the reported use of various post-hoc test statistical analyses, the observed frequencies of item construction errors, the observed frequencies of test format construction errors, and the observed cognitive functioning levels of test items on the teacher-made tests varied significantly by teacher grade level assignment (elementary and secondary) and by teachers' subject area specialization. Additionally, but with less consistency, the observed cognitive functioning levels and the observed frequencies of test item construction errors found on the teacher-made tests varied by years of teacher experience and by school setting (urban, rural, and suburban).

Most teachers (at least 54%) indicated that they never or rarely calculate means or standard deviations, complete item analyses procedures, or estimate the reliability of their teacher-made tests. On the other hand these teachers reported that they frequently prepared and gave many formal teacher-made tests during a typical school year. They reported extensive use of problem, multiple-choice, completion, and matching item types but less use of essay and true-false items. As Gullickson (1984) also reported, most teachers reported scheduling at least one formal test about every two weeks or more frequently in a typical class. The average number of formal tests scheduled by this

sample of classroom teachers in a typical school year was 54.1. Of the total number of items used in a school year, the teachers reported that approximately one of each four items was of the problem type, one in five was either a multiple-choice or completion type item, and only about one in ten items was either a true-false or matching type item.

In accord with the findings of Gullickson and Ellwein (1985) and Gullickson (1986), comparatively very few of these teachers reported regular use of post-hoc statistical procedures (e.g., computing reliability, means, standard deviations, etc.) on the results of their teacher-made tests. Further, and as Gullickson and Ellwein found, teacher responses to the items dealing with statistical procedures appeared to be somewhat inconsistent as many teachers in both studies reported completing estimates of test reliability but calculating means and standard deviations to a much lesser extent; most of us would assume the latter would typically be necessary before performing the former.

As Fleming and Chambers (1983) found the teacher-made tests analyzed in this study contained predominantly items functioning at the knowledge level (72%) and with the items measuring beyond the comprehension level being found almost exclusively on the math and science tests. Multiple-choice, matching, and short response type items were most frequently used on the teacher-made tests with essay item types being by far the least frequently

used. Relative to the frequency of item construction errors found on the teacher-made tests, the matching exercises were found to be by far the most error prone followed by completion, essay, and true-false item type exercises. The most frequently identified test format errors were absence of directions, unclear answering procedures, and test items not being numbered consecutively throughout the test.

The grade level classification (elementary and secondary) was found to be related in several instances to teachers' testing practices. Differences between elementary and secondary teacher responses were noted for five of the seven survey items devoted to testing practices, and the secondary teachers constructed proportionately more higher cognitive level test items and fewer knowledge level items than did the elementary grade teachers. The secondary teachers appeared to spend more time or emphasis on their testing as suggested by their reports of more frequent calculation of means and standard deviations, more frequent tests per course, more frequent use of item analysis procedures, and the personal construction of a larger proportion of the items used on their tests. Relative to the use of various item types, the secondary teachers reported relatively more use of essay and problem type items (which are often considered more appropriate for older students) and less frequent use of completion and multiple-choice items than did the elementary level teachers;

whereas the reported relative use of matching and true-false items did not differ between the elementary and secondary teachers.

The teachers' subject area classification resulted in the identification of teacher response differences for five of the seven survey items concerning testing practices, and each of the subject area classifications of the teacher-made tests differed significantly from the four others in terms of item cognitive level functioning with English teachers constructing proportionately the most comprehension level items, math teachers the most higher levels, and the social studies teachers proportionately the most knowledge level items. Many of these differences were caused by or associated with the math teachers (perhaps due to the relative uniqueness of the content in this subject area). The math teachers reported more use of problem type test items and less use of all other item types as compared to one or more of the other four groups of teachers, namely, business, science, English and social studies. A second but smaller group of differences was associated with the social studies teachers. These teachers reported less frequent calculation of means and standard deviations and less frequent use of item analysis procedures, but they reported more frequent writing of their own test items and more frequent use of completion and essay item types than did one or more of the other

teacher groups. The business education and the science teachers, like the math teachers, reported more frequent use of problem-type items than did the English and social studies teachers; the English teachers reported the most frequent use of essay-type items; and the business education and social studies teachers reported relatively more frequent use of true-false items than did the other teachers.

The analyses of the cognitive functioning levels of the teacher-made test items also revealed differences between teachers grouped by years of teaching experience and school setting. The less experienced teachers (1 to 3 years as compared to 4 to 6 or 7-10 years) wrote proportionately fewer knowledge level items and more comprehension level items than did the more experienced teachers, and the teachers employed in rural school settings as compared to the urban and suburban schools constructed proportionately fewer knowledge level items and more comprehension level test items.

In terms of possible implications or recommendations from the data gathered and analyzed from this sample of teachers the following are offered: a) Teachers are not convinced of the value of statistical procedures (or at least do not choose to use them) in improving and analyzing their teacher-made tests to the extent that measurement textbooks and professors emphasize these procedures. b) Teachers and their students expend considerable

effort and time in testing as indicated by the high frequency of tests teachers schedule in a typical school year. c) Inservice training should be provided periodically for teachers as it was found that teachers' testing practices and preferences did not change with additional years of teaching experience nor did the teacher-made tests of the more experienced teachers possess fewer errors. In fact, this study and Billeh (1974) found that more experienced teachers use even a greater proportion of knowledge level items than do the less experienced teachers. Further, many types of errors found on the teacher-made tests appeared to be of the nontechnical variety which could be addressed in relatively informal training settings. For example, the results of this study as well as that of Flemming and Chambers (1983) indicate that directions are absent in approximately one-fourth to one-third of teacher-made tests. d) Teachers need skills and/or encouragement to construct more higher cognitive level type test items. This study and others (Black, 1980; Flemming and Chambers, 1983) have indicated that most teacher-made test items measure primarily at the knowledge level, application level items are primarily limited to the math type subject area tests, and almost no items are found to measure at the analysis, synthesis, and evaluation levels. e) Most teachers (it was found in this study that approximately 80% of the teachers write one-half or more of the items used on their tests) write most of the test items used

on their tests which, along with the high frequency of item construction errors found on teacher tests, suggest a need for more preservice and inservice training emphasis on item writing skill development. f) Test writing skills appear to vary by subject area specialization with social studies teachers displaying a much lower level of skill development compared to other areas of specialization. g) Matching exercises need be given extra attention in training sessions as they are among the most frequently used item types and are the most error prone item type. h) Most teachers use a variety of item types; however, there appears to be a clear preference for differing item types among the subject area specializations; thus some attention in training sessions need be given to the subject field of the participants.

5/5

References

Black, T. R. (1980). An analysis of levels of thinking in Nigerian science teachers' examinations. Journal of Research in Science Teaching, 17, 301-306.

Billeh, V. Y. (1974). An analysis of teacher-made test items in light of the taxonomic objectives of education. Science Education, 58, 313-319.

Bloom, B. S., et al. (1956). Taxonomy of Educational Objectives: Handbook I, Cognitive Domain. New York: D. McKay.

Dwyer, C. A. (1982). Achievement testing. In H. E. Mitzel (Ed.), Encyclopedia of Educational Research, (4th ed., Vol. 1, pp. 13-22). New York: The Free Press.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. New Directions for Testing and Measurement, 19, 29-38.

Gullickson, A. R. (1984). Teacher perspectives of their instructional use of tests. Journal of Educational Research, 77, 244-248.

Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. Journal of Educational Measurement, 23, 347-354.

Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. Educational Measurement: Issues and Practice, Spring, 15-18.

Lambert, R. F. (1980-81). Teacher attitudes on testing: A multiple perspective. College Board Review, 29-30, 13-14.

Rogers, B. G. (1985). Prospective teacher perceptions of how classroom teachers use evaluation methods: A qualitative research approach. Mid-western Educational Researcher, 613-20.

Stiggins, R. J. (1985). Improving assessment where it means the most: In the classroom. Educational Leadership, 43, 69-74.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. Journal of Educational Measurement, 22, 271-286.

5/5

Table 1

Teacher Responses to Testing Practices and Preferences Survey Items

	% Never	% Rarely	% Occasionally	% Nearly Always	% Always
1. How often do you:					
a) Calculate test means and standard deviations?	49	31	14	4	1
b) Estimate reliability for your tests?	35	25	23	12	3
c) Complete item analysis of your tests (check item difficulty, etc.) to determine which questions "worked"?	29	25	28	13	3

d) Give formal (major) tests (exclude spelling or other quizzes) in a typical subject or class?	% One or more Each Week	% One Every Two Weeks	% One Per Month	% 3 or 4 Per Semester	% 2 or Fewer Per Semester
	20	49	15	7	6

2. What proportion of the questions used on your formal tests in a typical school year have you written yourself?	% Very Few	% About 1/4	% About 1/2	% About 3/4	% Almost All
	14	8	19	20	37

3. Including all classes or subjects taught, approximately how many formal tests (exclude spelling and other quizzes) do you give during a typical school year? (Hint: Check your grade book.) Number = _____.

$\bar{X}$ = 54.1;

Range of number of tests per year	
10 or more 92%	50 or more 42%
20 or more 75%	60 or more 31%
30 or more 58%	100 or more 15%

4. Of all test questions you use in a typical school year what approximate percentage of the total are of the following types? (Your percents should add to 100.)

Problems	22%	Essay	11%
Multiple Choice	20%	True/False	10%
Completion	17%	Other	4%
Matching	15%		

Table 2

F-Ratios and Grade Level and Subject Area Means for Teachers' Testing Practices

Grade Level Assignment Means

Practice		Elem.	2cdary	Total	F Value	p
1.a	Calculate $\bar{X}$'s & SD's	1.58	1.89	1.77	8.67	.01
1.b	Calculate Reliability	2.12	2.31	2.23	1.94	.17
1.c	Do Item Analysis	2.20	2.46	2.36	3.84	.05
1.d	No. Tests Course*	2.45	2.14	2.26	6.51	.01
2	Write Own Items	2.66	4.12	3.57	96.87	.00
3	No. Tests Per Year	47.62	53.65	51.43	0.89	.35

Subject Area Specialization Means

Practice**	(1) Bus.	(2) Sci.	(3) Math	(4) Eng.	(5) Soc. St.	Total	F	p	Scheffe***
1.a	1.86	2.39	1.95	1.73	1.46	1.89	4.18	.01	5<2
1.b	2.42	2.51	2.44	2.03	2.00	2.31	1.48	.21	-----
1.c	2.84	2.39	2.58	2.17	2.09	2.46	2.99	.02	5<1
1.d*	2.09	2.00	2.10	2.77	1.71	2.12	6.58	.00	4>1,2,3, & 5
2.	3.67	4.33	4.05	4.13	4.50	4.10	3.18	.02	5>1
3.	66.09	47.28	53.68	45.39	48.35	53.29	0.83	.51	-----

* Lower numbers here indicate more frequent test scheduling ('1' = one or more each week to '5' = two or fewer per semester).

** See top section of this table for item descriptions.

*** Scheffe' post-hoc pair-wise mean comparisons alpha @ .10.

Table 3

F-Ratios and Grade Level and Subject Area Means for Item Types Used by the Teachers

Grade Level Assignment Means

Item Type	Elementary	Secondary	Total	F	p
Completion	18.97	15.33	16.75	3.00	.084
Matching	13.79	14.46	14.19	.19	.666
True/False	9.32	9.74	9.58	.13	.724
Multiple Choice	24.48	16.72	19.74	11.21	.001
Essay	7.33	13.31	10.98	10.06	.002
Problems	13.98	26.33	21.51	12.59	.001
Other	4.16	2.91	3.39	.55	.460
(N)	(122)	(191)	(313)		

Subject Area Specialization Means

Item Type	(1) Bus.	(2) Sci.	(3) Math	(4) Eng.	(5) Soc. St.	Total	F	p	Scheffe*
Completion	15.44	16.58	7.66	13.90	21.97	14.78	3.54	.008	5>3
Matching	14.38	20.33	3.41	15.57	19.56	14.09	11.29	.001	3<1,2,4,5
True/False	14.69	8.52	3.44	7.20	14.56	9.75	12.10	.001	3<1,5; 4<1,5; 2<1
Multiple Choice	17.00	26.36	3.17	23.63	18.75	16.98	13.10	.001	3<1,2,4,5
Essay	7.44	9.85	.32	29.87	21.06	12.39	21.93	.001	5>1,2,3; 4>1,2,3
Problem	26.47	15.48	78.76	1.17	1.25	27.66	106.55	.001	3>1,2,4,5; 1>4,5, 2>4,5
Other	2.38	2.12	3.24	5.33	2.65	3.07	.37	.832	
(N)	(45)	(33)	(41)	(30)	(32)	(181)			

* Scheffe' post-hoc pair-wise comparisons alpha @ .10.

Table 4

Indications of Differences Among Subject Fields and Percentages of Exercise Type Found on the Total Number of Tests

	Subject Fields						
Exercise Type	English	Science	Business	Social Studies	Math	x^{2**}	p
Interpretive	18*	28	21	10	0	11.43	.02
Short Response	54	50*	62	63	31	8.31	.08
Essay	32	6	10	23	0	17.95	.01
Matching	75	53	38	43	6	32.69	.001
True/False	43	36	55	50	6	19.59	.001
Multiple-Choice	43	64	28	53	3	31.13	.001
Completion	25	22	45	53	0	26.56	.001
Problems	0	25	38	0	97	89.11	.001

*These values are to be interpreted as 18% of all English tests contained interpretive exercises, 50% of the science tests contained short response exercises, etc.

**Goodness of fit chi-square values computed on frequencies not on percentages.

Table 5

A. Item Cognitive Level Demands by Item Type

Item Type	N	% Beyond Knowledge	Knowl.	Compr.	Applic.	Analysis	Synthesis	Eval.
Completion	549	2	540	9	0	0	0	0
Matching	1261	8	1159	102	0	0	0	0
True/False	935	20	751	175	0	9	0	0
Multiple-Choice	1317	15	1123	7	112	73	2	0
Essay	64	53	30	22	6	1	1	4
Problems	896	96	35	59	798	4	0	0
Interpretive	362	35	199	118	40	4	0	1
Short Response	1093	24	830	235	28	0	0	0
Unclassified	52	46	28	23	0	0	1	0
Totals	6529		4695	750	984	91	4	5
% Cognitive Levels			72%	11%	15%	1%	.001%	.001%

B. Percentages of Items Functioning at Three Cognitive Levels by Various Teacher Classifications

Subject Classifications	Knowledge	Comprehension	High Levels	Pair-Comparison	x^2	p
Experience Teaching						
1-3 years	69	17	14	--	57.16	.001
4-6 years	73	12	15			
7-10 years	73	10	17			
School Setting						
Urban	74	10	16	--	31.08	.001
Rural	69	15	16			
Suburban	74	11	15			
Grade Level						
Elementary	76	18	6	--	111.05	.001
Secondary	71	12	17			
Teaching Field						
English (E)	77	21	2	E = S	93.04	.001
Science (S)	80	11	9	E = M	1515.13	.001
Math (M)	7	14	79	E = B	49.75	.001
Business (B)	79	14	7	E = SS	50.49	.001
Social Studies (SS)	98	2	0	S = M	1246.95	.001
				S = B	467.06	.001
				S = SS	184.46	.001
				M = B	1310.58	.001
				M = SS	1672.59	.001
				B = SS	208.34	.001

Table 6

Test Construction Error Type Summary of Frequencies

	No. Items Reviewed	% Total Items Reviewed	No. of Exercises	No. Errors Present*	Mean Errors Per Exercise
A. Item Type Errors					
1. Matching	1261	19	78	496	6.4
2. Completion	549	8	48	106	2.2
3. Essay	64	1	22	34	1.5
4. True/False	935	14	69	71	1.0
5. Multiple-Choice	1317	20	65	53	.8
6. Short Response	1093	17	89	61	.7
7. Problems	896	14	54	26	.5
8. Interpretive Exercise	362	6	30	6	.2
9. Unclassified	52	1	6	-	-
Subtotals	6529	99	455	853	1.9

B. Test Format Errors	No. Tests** Errors Present	% of Total
1. Absence of directions	82	29
2. Answering procedures unclear	61	22
3. Items not consecutively numbered	47	17
4. Adequate margins	22	8
5. Answer space provided	21	7
6. Space between items	12	4
7. Nonindependent items	11	4
8. Different weighting of objective items	8	3
9. Items arrange most to least time demanding	7	2
10. Similar item types not grouped together	6	2
	281	100

*Each specific item type construction error was tallied only once if present in an exercise (i.e., an error may have occurred several times or once in an exercise but in either case only a single tally was used so that tests and exercises could be compared regardless of the number of individual items appearing in a test or exercise).

**There were only 175 individual tests but some tests had more than one error.

Table 7

Frequency and Nature of Item Construction Errors on Teacher-Made Tests

Construction Error	No. Exercises* With Error	% of Total Errors This Type
a. Completion Item Type:		
1. Not complete interrogative sentence	32	30
2. Blanks in statements	31	29
3. Textbook statements with words left out	18	17
4. More than single blank in statement	12	11
5. Question allows more than single answer	6	6
6. Blank number clue	4	4
7. Blank length clue	1	1
8. Requests trivia versus significant idea	1	1
9. Unstated degree of precision	1	1
10. Lengthy, unnecessary words or phrases	0	0
	106	100
b. True/False or Alternate Response		
1. Required to write response, time waste	20	28
2. Statements contain more than single idea	16	23
3. Negative statements used	15	21
4. Presence of specific determiner	8	11
5. Statement not question, give away item	6	8
6. Needless phrases present, too lengthy	4	6
7. Imprecise statement, not always true or false	1	2
8. Presence of length clue	1	1
9. Opinion not attributed to source	0	0
	71	100

*Each specific item type construction error was tallied only once if present in an exercise (i.e., an error may have occurred several times or once in an exercise but in either case only a single tally was used so that tests and exercises could be compared regardless of the number of individual items appearing in a test or exercise).

(table continues)

Construction Error	No. Exercises With Error	% of Total Errors This Type
c. Matching Item Type:		
1. Columns not titled	71	14
2. Use one, more than once, or not all not in directions to prevent elimination	69	14
3. Response column not ordered	60	12
4. Directions not specify basis for match	55	11
5. Answering procedure not specified	52	10
6. Elimination due to equal numbers	46	9
7. Column(s) exceed 10 items	39	8
8. Materials not homogeneous	38	8
9. Premise not to left side	37	7
10. Numbers not to left and letters to right	13	3
11. Exercise not contained on single page	7	2
12. Requires responses to be written out	6	1
13. Insufficient information in premises	3	1
	496	100
d. Multiple Choice		
1. Alternates not in column(s)	21	40
2. Incomplete stems	12	23
3. Negative words not underlined	9	17
4. All or none above not approximately used	5	9
5. Needless repetition in alternates	2	4
6. Presence of specific determines in alternates	2	4
7. Verbal associations between alternate and stem	1	2
8. Alternates overlap	1	1
9. Needless phrases used	0	0
10. Grammatical clues	0	0
11. Distractors implausible	0	0
12. Length clues	0	0
13. a and c, but not b, etc. used	0	0
	53	100

(table continues)

Construction Error	No. Exercises With Error	% of Total Errors This Type
e. Essay Exercises		
1. Response expectations clear, labeled, etc.	14	41
2. Scoring points not realistically limited	7	21
3. Optional questions provided	5	15
4. Restricted question not provided	3	9
5. Ambiguous words used	2	6
6. Opinion or feelings requested	2	6
7. Question limited to single listing response	1	2
	34	100
f. Problem Exercises		
1. Items not sample understanding concepts, only calculations	20	77
2. Not range of easy to difficult problems	3	12
3. Degree of accuracy not requested	2	8
4. Nonindependent items	1	4
5. Use of objective items when calculation preferable	0	0
	26	100
g. Interpretive Exercises		
1. Objective response form not used	6	100
2. Can be answered without data presented	0	0
3. Errors present in response form	0	0
4. Data presented unclear	0	0
	6	100
h. Short Response		
1. Item requires listing, recall only	51	84
2. Response expectations ambiguous, not specified	7	11
3. Unrealistically high scoring values assigned	3	5
	61	100